Towards Attuned AI: Integrating Care Ethics in Large Language Model Development and Alignment

Rayane El Masri*
Generative AI Lab
Queensland University of Technology
Brisbane, Australia
rayane.elmasri@hdr.qut.edu.au
ADMS+ Centre Profile

Aaron Snoswell
Generative AI Lab
Queensland University of Technology
Brisbane, Australia
a.snoswell@qut.edu.au
ADMS+ Centre Profile

Abstract

How can the Ethics of Care (EoC) inform the development and value alignment of large language models (LLMs)? This paper proposes to investigate how a Care ethics framework emphasizing relationality, attention to particularities, and contextual moral reasoning, can reshape existing approaches to aligning LLMs with human values. Mainstream AI alignment often draws on deontological or utilitarian principles, yet these frameworks can overlook the situated, affective, and power-sensitive aspects of moral life that Care ethics foregrounds. In this paper we present two arguments for integrating EoC into LLM development practices. First, we argue that LLMs often rely on overly generalized reasoning, which contributes to various down-stream harms, including issues of bias. Second, we critique methods like RLHF and RLAIF for embedding narrow normative assumptions that neglect emotional and relational dimensions of human values. We argue that adapting LLM fine-tuning or alignment practices to incorporate Ethics of Care considerations may help address these issues, potentially laying the groundwork for better forms of LLM generalization and providing a pathway for more context sensitive alignment of LLMs in care-relevant areas such as mental health, education, and social services.

1 Background and Introduction

1.1 Introduction

In this paper, we explore how a care ethics framework that emphasizes relationality, attention to particularities, and contextual moral reasoning can reshape the way we approach aligning LLMs with human values. While mainstream AI alignment methods often draw from utilitarian or deontological principles, these frameworks frequently overlook the situated, affective, and power-sensitive dimensions of moral life that EoC brings to the foreground. We argue that integrating care ethics into the design and value alignment (VA) processes of LLMs offers a richer and more contextually grounded normative foundation for generative AI systems. To support this claim, we develop two main arguments. In the first argument, we examine the problem of generalization in LLMs and argue that these models often fail to respond appropriately to individual needs and social contexts because they apply overly generalized patterns of reasoning. This failure makes them ethically inadequate in emotionally sensitive domains. In the second argument, we critique the current normative commitments embedded in popular VA methods such as Reinforcement Learning from Human Feedback

*

Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS 2024.

(RLHF) and Reinforcement Learning from AI Feedback (RLAIF). These approaches largely reflect utilitarian and justice-based theories, which tend to abstract away from relational and emotional dimensions of moral experience. We argue that these methods are structurally limited and that EoC may offer a more appropriate framework for some care-relevant domains such as mental health, education, and social services.

1.2 Introducing the Ethics of Care

The ethics of care, rooted in feminist moral philosophy, contrasts sharply with justice-based theories that have historically dominated Western culture by focusing on moral responsiveness and the lived realities of interdependence. Carol Gilligan Gilligan [1993] introduced EoC as a critique of Lawrence Kohlberg's (1981) justice-centered model of moral development, which prioritized rights and rule-based reasoning [Gilligan, 1993]. Gilligan demonstrated that this model, based largely on research with male subjects, overlooked care-based reasoning more commonly expressed by women, and mischaracterized it as morally immature. She argued instead that care and empathy represent a distinct and equally valid moral choice. This insight laid the groundwork for further contributions by Nel Noddings Noddings [1984], Virginia Held Held [2005b], and Joan Tronto Tronto [1998], all of whom emphasized the ethical significance of attentiveness, responsiveness, and the contextual nature of human relationships. Unlike frameworks grounded in impartiality and universality, EoC focuses on particular needs and vulnerabilities, making it particularly well-suited for LLMs deployed in domains and contexts where providing care is a fundamental aspect of the interaction, such as mental health support.

Some scholars have already called for integrating EoC into AI ethics. For example, Cohn [2020] proposes care ethics as a human-centered framework for AI development, critiquing efficiency-driven models that ignore vulnerability and social inequality [Cohn, 2020]. He argues that care should be habitual and grounded in relational understanding. However, his proposal lacks concrete implementation strategies, and is framed with respect to AI development generally, not LLMs specifically. More recently, Villegas-Galaviz and Martin [2024] warn that the increasing use of AI in decision-making is creating new forms of moral distance. They distinguish between proximity distance, caused by the removal of face-to-face interaction, and bureaucratic distance, arising from hierarchical structures and abstraction [Villegas-Galaviz and Martin, 2024]. These forms of distancing, they argue, result in ethical neglect. They argue for care ethics as a potential remedy, but they do not provide a roadmap for operationalizing it in LLM development or alignment.

Addressing this gap, we argue that EoC holds immense potential for addressing the moral limitations of current AI systems. Its emphasis on empathy, moral attentiveness, and responsiveness provides a much-needed corrective to dominant utility-maximizing approaches, especially in applications like mental health support or eldercare [Cohn, 2020, Villegas-Galaviz and Martin, 2024]. By resisting abstraction and centering relationality, EoC challenges the moral distancing introduced by automation and algorithmic decision-making [Gilligan, 1993, Held, 2005a]. Its compatibility with non-Western ethical traditions such as Ubuntu, "I am because we are", also makes it a powerful decolonial alternative to dominant AI ethics frameworks [Amugongo et al., 2023].Although care ethics presents challenges for AI implementation. Its contextual and relational nature resists formalization and scalability [Weinberger, 2024] and its lack of procedural structure may hinder transparency and consistency in automated systems [Gabriel, 2020, Gabriel and Ghazavi, 2021]. These features may complicate the direct translation of care principles into algorithmic specifications for Large Language Models. However, they are not reasons to exclude care ethics but to adapt our design and alignment strategies to accommodate its insights. Through our two arguments, one from the dissonance of LLM's generalizations and another from the shortcomings of current normative frameworks, we aim to show why EoC is not only a relevant but necessary ethical foundation for the future of responsible LLMs' development.

2 Arguments for the integration of Ethics of Care in LLMs development

An argument from the dissonance of LLM generalizations Feed-forward Neural Networks (FFNNs) are known to be universal function approximators capable of generalizing to new instances within the bounds of their training data. This theoretical capacity was established early on through work in Statistical Learning Theory and the development of back-propagation algorithms [Plaut et al.,

1986, Rumelhart et al., 1985, Vapnik, 2000]. However, a critical limitation arises when considering the real-world deployment of FFNNs, particularly in the context of Large Language Models. While the mathematical theories accurately predict performance degradation on data significantly deviating from the training distribution, the core issue lies in the very concept of a well-defined 'training data set distribution' being representative of the dynamic and complex real world.

As Marcus [1998] observed in 1998, empirical evidence shows that neural networks often falter on out-of-distribution data, leading to unreliable or even harmful outputs [Marcus, 1998]. This isn't merely a failure to generalize beyond the training data-set, it underscores a fundamental problem. The 'real world' where LLMs operate is rarely as neatly categorized as a static training distribution might suggest. Human language and the world it reflects are inherently messy, with users and situations frequently existing in the 'margins' or as 'edge cases' that don't perfectly align with the training data. Consequently, the performance of FFNNs in these seemingly 'edge' scenarios becomes a crucial consideration, highlighting a significant gap between theoretical capabilities within controlled distributions and the unpredictable nature of real-world applications.

This limitation extends directly to modern Large Language Models (LLMs), which are, at their core, massively scaled-up versions of FFNNs trained using sophisticated forms of back-propagation [Rumelhart et al., 1985]. As such, they are vulnerable to the same generalization problems. We see this in practice through empirical evidence that LLMs frequently exhibit representational biases and produce context-insensitive or inappropriate outputs, a problem extensively documented in recent work [Navigli et al., 2023]. A notable example of political bias in Large Language Models (LLMs) was documented in a study published in the Journal of Economic Behavior and Organization in early 2025. The research confirmed a left-leaning bias in ChatGPT, noting its tendency to produce content aligned with left-wing values and occasional restrictions on right-leaning themes [Motoki et al., 2025].

If LLMs in practice rely on complex rules to apply knowledge to new contexts, rules which often fail to generalize appropriately in practice, then continuing to build systems that depend on such generalization is ethically and functionally problematic. Illustrative of the inherent limitations of relying on abstract generalizations in Large Language Models (LLMs) is their propensity to replicate harmful societal biases across critical domains. Despite aspirations of neutrality, these models frequently reflect and amplify prejudices present in their training data and human feedback mechanisms, yielding tangible real-world consequences. To illustrate the breadth and systemic nature of this issue, consider the following examples across mental health, gender, and social services.

In the realm of mental health, research has shown therapy chatbots exhibiting discriminatory responses based on users' conditions, perpetuating harmful stigmas [Harrison Dupré, 2025]. Regarding gender, LLMs have been found to generate narratives that reinforce traditional stereotypes in storytelling, demonstrating a clear gender bias in creative outputs [UCL, 2024]. Furthermore, in the context of social services, attempts to implement AI for fairer benefit assessments have resulted in the reproduction of existing human biases, highlighting the challenge of achieving equitable outcomes through purely algorithmic means [Guo et al., 2024]).

These instances demonstrate how LLMs, despite their technical sophistication, can inadvertently entrench societal inequities in care-relevant applications. This bias is not solely a product of flawed training data but also arises from the subjective values embedded in human and even AI feedback during alignment processes. Consequently, relying on dominant moral frameworks that prioritize abstract utility or generalized fairness proves inadequate for addressing these nuanced biases. The Ethics of Care, with its focus on attentiveness to individual needs and relational context, offers a crucial alternative for mitigating the harms of generalization and fostering more equitable and sensitive AI systems. The issue is not that these models operate on "simple" or "complicated" rules, but rather that their internal logic frequently fails to track the particularities that matter in real-world contexts, no matter how intricate it is. Instead, we should aim for systems that are attuned to the particularities of individuals and their contexts of LLM usage, rather than imposing broad generalizations that may misrepresent, exclude, or even harm them.

The Ethics of Care (EoC) explicitly promotes this kind of sensitivity to individual circumstances as a core ethical principle. Care theorists emphasize relationality, context, and attentiveness to vulnerability and dependence, which directly contrasts with the detached and abstract logic of generalizing systems. Re-thinking the training and design of LLMs through the lens of EoC would entail prioritizing individual and contextual particularities rather than optimizing for statistical

generalization. This shift could help mitigate the harmful consequences of misapplied generalizations and align AI systems more closely with the nuanced moral demands of human interaction.

An argument from the Shortcomings of Current Normative Commitments of VA methods in Care-Related Contexts Current Value Alignment (VA) methods, particularly Reinforcement Learning from Human Feedback (RLHF) and its more recent variant Reinforcement Learning from AI Feedback (RLAIF), are grounded in specific normative frameworks, primarily utilitarianism and justice-based theories such as Contractualism and Rawlsian justice. These approaches optimize behaviour either by aggregating human preferences (in RLHF), or by aligning model outputs with predefined fairness metrics or constitutional principles (in RLAIF), thus reflecting a commitment to maximizing overall utility or ensuring procedural fairness [Sola, 2023]. These moral commitments, however, are not value-neutral. Utilitarianism assumes that a morally right action is the one that maximizes utility, often measured as human satisfaction or preference fulfillment. Justice-based theories, such as the theory of justice [Rawls, 1971] or broader forms of Contractualism, prioritize fairness and equality through rules or distributions, aiming for consistency and impartiality. Yet, these normative theories exhibit important limitations when applied to AI systems operating in care-relevant contexts. Utilitarian approaches may justify sacrificing individual well-being for the sake of aggregated good, which is ethically troubling in contexts such as mental health or healthcare, where individual vulnerability and situational complexity demand more than mere utility maximization [Sola, 2023]. Likewise, justice-based approaches may overemphasize principles such as truthfulness or impartiality, and fail to account for the emotional, relational, and context-sensitive needs of individuals in morally complex or care-relevant situations [Gilligan, 1993].

Fundamentally, RLHF and RLAIF are built to maximize a reward signal, a proxy for desirable behaviour based on human or AI feedback. This feedback often embodies utilitarian assumptions: that optimizing the aggregate of preference scores leads to morally appropriate outcomes. But this moral calculus does not reflect how humans actually make ethical decisions. Research in moral psychology, particularly by [Gilligan, 1993], has shown that moral reasoning is not purely abstract or rule-based but is often relational and context-sensitive. Humans consider the specific effects of their actions on others, not just whether a rule was followed or a utility score maximized. This disconnect becomes especially critical as LLM-based systems are increasingly deployed in domains where care, empathy, and responsiveness are central. Examples include healthcare, education, eldercare, and mental health support, these fields in which the needs of individuals are unique, emotionally laden, and highly context-dependent [Beauchamp, 2006]. In such settings, the current VA frameworks, grounded in rule-following and aggregation, prove inadequate.

Claims of adherence to deontological ethical frameworks are common in the contemporary discourse on value alignment (VA) of Large Language Models (LLMs). A notable instance is Anthropic's Constitutional AI, which explicitly adopts a deontological or contractual approach. At its core is a detailed "constitution," a set of rules inspired by documents like the Universal Declaration of Human Rights and professional ethical codes [Sanwal, 2025]. This "constitution" is designed to promote three deontic commitments: harmlessness, helpfulness, and honesty, often referred to as the three H's [Sanwal, 2025]. On the surface, this suggests a clear ethical orientation grounded in deontological or contractual reasoning. However, the implementation of this model reveals a different logic. After a supervised fine-tuning phase in which the model critiques its own responses using the constitution as a reference, the training process transitions to reinforcement learning. At this stage, an AI preference model trained to align with the constitution provides reward signals, and the system is optimized to maximize these signals [Bai et al., 2022]. In practice, while the constitution declares a duty-based moral stance, "do no harm," "be helpful," "be honest", the underlying alignment mechanism remains reward-driven: responses are ultimately chosen based on which maximizes the preference-model's score [Bai et al., 2022]. Thus, Anthropic's framework presents a deontological front, but its mechanics operate on utilitarian principles. This reliance on singular deontological frameworks, or their utilitarian operationalization, exhibits inherent limitations when confronted with the complex, emotionally salient, and relationally embedded contexts in which LLMs are increasingly deployed, particularly within care-sensitive domains such as mental health, education, and social services. Abstract, generalized principles, or the pursuit of aggregated preference maximization, often fail to adequately address the specific vulnerabilities, individual needs, and contextual particularities inherent in these domains. The ethics of care (EoC) offers a corrective to these limitations. Instead of prioritizing universal rules or aggregate utility, EoC foregrounds attentiveness to particularities, responsiveness to vulnerabilities, and the understanding that moral reasoning is often iteratively

constructed within relationships. Integrating the principles of EoC into the VA pipeline, whether through mechanisms that modify Reinforcement Learning from Human Feedback (RLHF), Reinforcement Learning from AI Feedback (RLAIF), or constitutional frameworks, has the potential to yield more comprehensive and desirable behaviours from LLMs. What might this look like in practice? Incorporating care ethics into RLHF or RLAIF could involve modifying how feedback is interpreted and how behaviour is reinforced, moving away from generic reward maximization and toward recognition of individual and contextual moral significance. For example, in the training of an AI mental health chatbot, standard RLHF might reward responses that align with broadly preferred traits like honesty or politeness. However, a care ethics-informed chatbot would go further by weighing feedback according to the user's current unique emotional state, relational cues, and expressed needs. A distressed user expressing suicidal thoughts might receive an AI response that prioritizes empathy, the user's safety, and relational responsibility over truthfulness. Rather than reinforcing a response simply because it is rated highly in general by annotators, the model would be trained to identify morally salient features such as signs of vulnerability, dependency, or self-harm in this particular interaction and then be allowed to adjust its behaviour accordingly. As current VA methods remain systematically limited by their reliance on abstraction and generalization, care ethics offers a complementary framework. It helps fill the moral gaps left by traditional theories and provides more appropriate guidance for AI systems operating in human-centric and care-sensitive environments [Maio, 2018]. Thus, the case of Constitutional AI highlights the inadequacy of purely abstract approaches to the value alignment problem. As we have seen, although in principle, Anthropic adopted a deontological approach to value alignment, the technical implementation could only be utilitarian. Therefore, we argue that it is necessary to develop a value alignment method tailored to care-relevant domains and contexts inspired by the ethics of care. When it comes to the technical implementation of this VA method, current methods such as RLHF and RLAIF can be modified to incorporate care-based principles. However, the comprehensive elaboration of the full technical methodology is beyond the scope of this paper. More research still needs to be done. In this paper, we have argued that the development of AI systems genuinely aligned with human values, especially within contexts involving human vulnerabilities, necessitates the adoption of ethical frameworks such as care ethics, which are deeply rooted in the moral realities of lived experiences.

3 Conclusion

Integrating the ethics of care into the development of LLMs and value alignment processes offers a promising path forward that addresses the failures of generalization and the limitations of prevailing normative frameworks. A care-based VA approach would train systems not merely to be reward-sensitive, but to be morally attentive, i.e. responsive to the emotional, social, and contextual needs of users. Rather than optimizing for abstract utility or procedural fairness, care ethics would guide alignment processes like RLAIF to consider the specific relational context and potential harms to vulnerable individuals and communities. It would also inform the work of human labelers, encouraging them to reflect not only on individual preferences, but on the broader relational impacts of AI responses, especially for marginalized or care-dependent populations. Constitutionally, care ethics could provide a richer moral foundation by promoting the well-being of both caregivers and care-receivers within interdependent social networks. By embedding this moral orientation, LLM's chatbots would be more likely to resist sycophantic behaviour, minimize hallucinations, and engage users with a deeper sense of responsibility and trust. By offering a context-sensitive and relationally grounded framework, care ethics better reflects how people actually want to be treated in morally significant interactions with AI. While the Ethics of Care (EoC) offers a promising alternative to dominant value alignment (VA) frameworks, integrating it into the development of large language models (LLMs) poses several challenges. EoC's emphasis on contextual-relational reasoning resists formalization, making it difficult to operationalize within current machine learning architectures that rely on abstraction and generalization. Unlike rule-based or utilitarian models, care ethics does not offer easily codifiable principles, which complicates its translation into algorithmic procedures. Also, implementing EoC requires LLMs to recognize and respond to complex emotional and social cues, a capacity that remains technically underdeveloped. Current models lack genuine understanding of vulnerability, dependency, and relational dynamics, raising concerns about their ability to meaningfully provide care-based reasoning. These limitations highlight the need for more research and the importance of interdisciplinary collaborations between researchers and cautious implementation strategies when considering care ethics in AI design.

Acknowledgements

We would like to gratefully acknowledge the support of the Queensland University of Technology (QUT), particularly the Digital Media Research Centre (DMRC), the Generative AI Lab, and the ADMS+ Centre. This research was made possible through the support of the Generative AI Lab Scholarship.

References

Lameck Mbangula Amugongo, Nicola J. Bidwell, and Caitlin C. Corrigan. Invigorating Ubuntu Ethics in AI for healthcare: Enabling equitable care. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 583–592, Chicago IL USA, June 2023. ACM. ISBN 979-8-4007-0192-4. doi: 10.1145/3593013.3594024.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback, December 2022.

Tom L. Beauchamp. The 'Four Principles' Approach to Health Care Ethics. In Richard E. Ashcroft, Angus Dawson, Heather Draper, and John R. McMillan, editors, *Principles of Health Care Ethics*, pages 3–10. Wiley, 1 edition, June 2006. ISBN 978-0-470-02713-4 978-0-470-51054-4. doi: 10.1002/9780470510544.ch1.

Jonathan Cohn. In A Different Code: Artificial Intelligence and The Ethics of Care. *The International Review of Information Ethics*, 28, June 2020. ISSN 1614-1687. doi: 10.29173/irie383.

Iason Gabriel. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3):411–437, September 2020. ISSN 1572-8641. doi: 10.1007/s11023-020-09539-2.

Iason Gabriel and Vafa Ghazavi. The Challenge of Value Alignment: From Fairer Algorithms to AI Safety, January 2021.

Carol Gilligan. *In a Different Voice: Psychological Theory and Women's Development*. Harvard University Press, 1993. ISBN 978-0-674-44543-7. doi: 10.2307/j.ctvjk2wr9.

Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. Bias in Large Language Models: Origin, Evaluation, and Mitigation, November 2024.

Maggie Harrison Dupré. Stanford Research Finds That "Therapist" Chatbots Are Encouraging Users' Schizophrenic Delusions and Suicidal Thoughts. https://futurism.com/stanford-therapist-chatbots-encouraging-delusions, June 2025.

Virginia Held. Care and the Extension of Markets. In Virginia Held, editor, *The Ethics of Care: Personal, Political, and Global*, page 0. Oxford University Press, December 2005a. ISBN 978-0-19-518099-2. doi: 10.1093/0195180992.003.0008.

Virginia Held. Care as Practice and Value. In Virginia Held, editor, *The Ethics of Care: Personal, Political, and Global*, page 0. Oxford University Press, December 2005b. ISBN 978-0-19-518099-2. doi: 10.1093/0195180992.003.0003.

Giovanni Maio. Fundamentals of an Ethics of Care. In Franziska Krause and Joachim Boldt, editors, *Care in Healthcare*, pages 51–63. Springer International Publishing, Cham, 2018. ISBN 978-3-319-61290-4 978-3-319-61291-1. doi: 10.1007/978-3-319-61291-1_4.

Gary F. Marcus. Rethinking Eliminative Connectionism. *Cognitive Psychology*, 37(3):243–282, December 1998. ISSN 0010-0285. doi: 10.1006/cogp.1998.0694.

Fabio Y. S. Motoki, Valdemar Pinho Neto, and Victor Rangel. Assessing political bias and value misalignment in generative artificial intelligence. *Journal of Economic Behavior & Organization*, 234:106904, June 2025. ISSN 0167-2681. doi: 10.1016/j.jebo.2025.106904.

Roberto Navigli, Simone Conia, and Björn Ross. Biases in Large Language Models: Origins, Inventory, and Discussion. *J. Data and Information Quality*, 15(2):10:1–10:21, June 2023. ISSN 1936-1955. doi: 10.1145/3597307.

Nel Noddings. *Caring: A Feminine Approach to Ethics and Moral Education*. University of California Press, 1984.

David C Plaut, Steven J Nowlan, and Geoffrey E Hinton. Experiments on Learning by Back Propagation. June 1986.

John Rawls. *A Theory of Justice*. Harvard University Press, 1971. ISBN 978-0-674-00078-1.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning Internal Representations by Error Propagation. 1985.

Manish Sanwal. Constitutional AI: An Expanded Overview of Anthropic's Alignment approaches. May 2025. doi: 10.5281/zenodo.15331063.

Andrew Sola. Utilitarianism and Consequentialist Ethics: Framing the Greater Good. In Andrew Sola, editor, *Ethics and Pandemics: Interdisciplinary Perspectives on COVID-19 and Future Pandemics*, pages 61–83. Springer Nature Switzerland, Cham, 2023. ISBN 978-3-031-33207-4. doi: 10.1007/978-3-031-33207-4_4.

Joan C. Tronto. An Ethic of Care. *Generations Journal*, 22(3):15–20, 1998. ISSN 07387806.

UCL. Large Language Models generate biased content, warn researchers. https://www.ucl.ac.uk/news/2024/apr/large-language-models-generate-biased-content-warn-researchers, April 2024.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, NY, 2000. ISBN 978-1-4419-3160-3 978-1-4757-3264-1. doi: 10.1007/978-1-4757-3264-1.

Carolina Villegas-Galaviz and Kirsten Martin. Moral distance, AI, and the ethics of care. *AI & SOCIETY*, 39(4):1695–1706, August 2024. ISSN 1435-5655. doi: 10.1007/s00146-023-01642-z.

David Weinberger. The Rise of Particulars: AI and the Ethics of Care. *Philosophies*, 9(1):26, February 2024. ISSN 2409-9287. doi: 10.3390/philosophies9010026.